DATA NOTE

Open Access

The draft genomes of *Crassostrea gasar* and *Crassostrea rhizophorae*: key resources for leveraging oyster cultivation in the Southwest Atlantic

Nicholas Costa Barroso Lima[1], Luiz Gonzaga Paula de Almeida[2], Afonso Celso Dias Bainy[3], Alexandra Lehmkuhl Gerber[2], Ana Paula de Campos Guimarães[2], Antonio Mateo Solé-Cava[4], Claudio Manoel Rodrigues de Melo[2], Cristiano Lazoski[5], Flávia Lucena Zacchi[6], Frederico Henning[4], Leticia Maria Monteiro Soares[4], Rafaela Guilherme Soares[4] and Ana Tereza Ribeiro Vasconcelos[2*]

Abstract

Objectives The two oyster species studied hold considerable economic importance for artisanal harvest (*Crassostrea rhizophorae*) and aquaculture (*Crassostrea gasar*). Their draft genomes will play an important role in the application of genomic methods such as RNAseq, population-based genomic scans aiming at addressing expression responses to pollution stress, adaptation to salinity and temperature variation, and will also permit investigating the genetic bases and enable marker-assisted selection of economically important traits like shell and mantle coloration and resistance to temperature and disease.

Data description The draft assembly size of *Crassostrea gasar* is 506 Mbp, and of *Crassostrea rhizophorae* is 584 Mbp with scaffolds N50 of 11,3 Mbp and 4,9 Mbp, respectively. The general masked bases by RepeatMasker in both genomes were highly similar using different datasets. The masked bases varied from 9.41% in *C. gasar* to 10.05% in *C. rhizophorae* and 42.85% in *C. gasar* to 44.44% in *C. rhizophorae* using Dfam and RepeatModeler datasets, respectively. Functional annotation with eggNog resulted in 34,693 annotated proteins in *C. rhizophorae* and 26,328 in *C. gasar*. BUSCO analysis shows that almost 99% of genes (5,295) are complete in relation to the mollusk orthologous genes dataset (mollusca_odb10).

Keywords *Mollusca*, *Crassostrea*, *Gasar*, *Rhizophorae*, Genome

*Correspondence:
Ana Tereza Ribeiro Vasconcelos
atrv@lncc.br

[1] Departamento de Bioquímica e Biologia Molecular, Centro de Ciências, Universidade Federal do Ceará (UFC), Fortaleza, CE 60020-181, Brasil

[2] Laboratório de Bioinformática, Laboratório Nacional de Computação Científica, Av. Getúlio Vargas, 333 - Quitandinha, Petrópolis, RJ 25651-076, Brasil

[3] Laboratório de Biomarcadores de Contaminação Aquática e Imunoquímica, Departamento de Bioquímica, Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC 88037-000, Brasil

[4] Centro Nacional para a Identificação Molecular do Pescado (CENIMP), Departamento de Genética, Instituto de Biologia, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ 21941-590, Brasil

[5] Laboratório de Biodiversidade Genômica (LABIG), Departamento de Genética, Instituto de Biologia, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ 21941-902, Brasil

[6] Laboratório de Moluscos Marinhos (LMM), Departamento de Aquicultura, Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC 88061-600, Brasil

Objective

The two oyster species whose draft genomes we publish here, *Crassostrea rhizophorae* and *Crassostrea gasar* hold considerable economic importance for artisanal harvest and aquaculture. Their draft genomes will play an important role in answering different questions. *C. rhizophorae* grows across a wide range of environments despite varying degrees of environmental stress and is commonly used as a sentinel and bioindicator species in environmental monitoring studies. Using RNAseq, population genomic analysis, and RAD markers, we are comparing *C. rhizophorae* oyster samples from heavily polluted and pristine areas in Rio de Janeiro, Paraná, and Santa Catarina States. This comparison aims to elucidate metabolic pathways, identify loci under selection, and gain insights into the adaptation mechanisms of these oysters to pollution, ultimately designing an effective biomonitoring system. Efficient use of reduced genomic representation methods for population genomics requires genome sequences to locate associated markers. *Crassostrea gasar* is particularly suitable for cultivation and exhibits traits of economic importance, such as shell and mantle coloration and resistance to temperature and salinity variations. These traits must be artificially selected to improve yield and market value. The genomes produced will help identify the genetic bases of these important traits. Through population genomics, transcriptomics, and forward-genetics, we can effectively assist in their artificial selection via Marker Assisted Selection (MAS) to improve aquaculture production of the species. Therefore, by making these data available, we aim to collaborate on genomics studies across oysters.

Data description

The specimens of *Crassostrea rhizophorae* used for PacBio CLR, PacBio HiFi, MinIon (Oxford Nanopore Technologies), and Illumina sequencing were sampled from natural outbred population at Praia da Boa Viagem (Niterói, RJ, Brazil), Praia da Caieira da Barra do Sul (Florianópolis, SC, Brazil) and Rio Bücheller (Florianópolis, SC, Brazil) (Reads summary in Table 1—Data Set 1 (Table 1)). The specimens of *Crassostrea gasar* used for PacBio CLR, HiFi, and Illumina sequencing originated from the stock maintained at the Laboratory of Marine Mollusk at the Federal University of Santa Catarina (UFSC) (Reads summary in Table 1—Data Set 1 (Table 1)). Specimens were dissected live for mantle tissues.

A schematic of the assembly, gene prediction, and annotation process for both genomes is shown in Fig. 1. We used the genomes and proteins of *C. angulata*, *C. gigas*, and *C. virginica* for comparison with *C. gasar* and *C. rhizophorae* draft assembly and predicted genes. Assembly was performed with Hifiasm v0.19.5 and ntLink v1.3.9 [1–4], with scaffold gap-filling done using GapFiller v1-11 [5]. After assembly and gap-filling, the final drafts were checked for completeness and basic assembly statistics using BUSCO v5.4.7 and Quast v5.2.0 [6, 7]. Repeat identification and masking for all five genomes was carried out with RepeatMasker v4.1.6, RepeatModeler v2.0.5, and Dfam v3.8 [8–10]. Gene prediction was performed with the BRAKER pipeline v3.0.3, employing AUGUSTUS and GeneMarker-ET based on RNA-seq [11]. Functional annotation of predicted genes used eggNOG v2.1.10 and Diamond v2.1.9 [12, 13].

The draft assembly size of *C. gasar* was 506 Mbp, and *C. rhizophorae* was 584 Mbp, with scaffold N50 sizes of 11.3 Mbp and 4.9 Mbp, respectively (Table 1—Data Set 1 (Table 2)). BUSCO analysis showed that nearly 99%

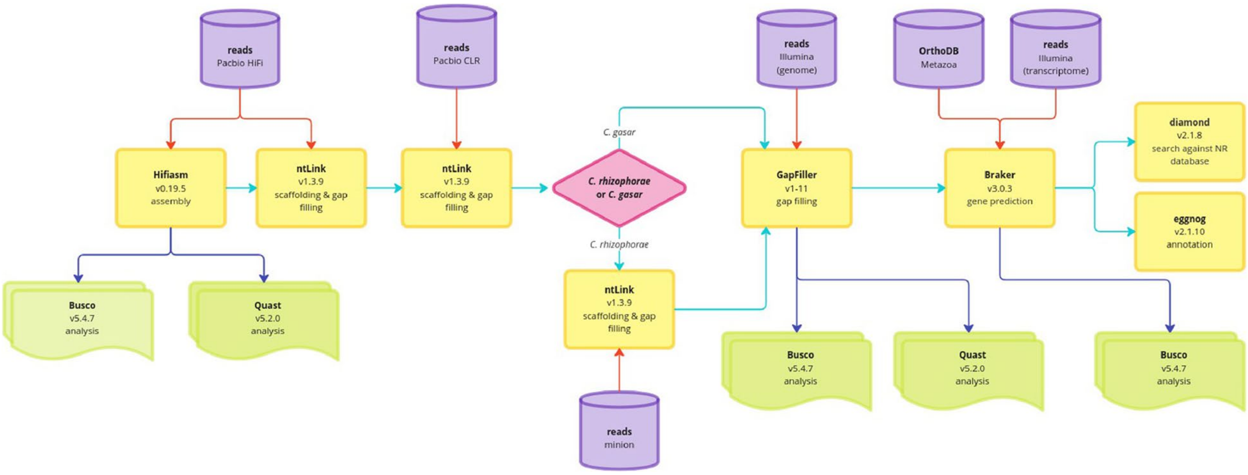

Fig. 1 Pipeline for the assembly and annotation of the draft genomes of C. gasar and C. rhizophorae

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Figure_1	portable document format (.pdf)	https://doi.org/10.5281/zenodo.12103998 [20]
Data File 2	gasar_annotation	general transfer format (.gtf)	https://doi.org/10.5281/zenodo.12103998 [20]
Data File 3	rhizophorae_annotation	general transfer format (.gtf)	https://doi.org/10.5281/zenodo.12103998 [20]
Data File 4	Crassostrea gasar draft genome	genbank format (.gbk)	https://identifiers.org/ncbi/nucleotide:JBEEQF000000000.1 [18]
Data File 5	Crassostrea rhizophorae draft genome	genbank format (.gbk)	https://identifiers.org/ncbi/nucleotide:JBEOLP000000000.1 [19]
Data set 1	Tables	spreadsheets (.xlsx)	https://doi.org/10.5281/zenodo.12103998 [20]

of genes (5,295) in the mollusk orthologous genes dataset (mollusca_odb10) are complete (Table 1—Data Set 1 (Table 3)). The number of repetitive sequences across all analyzed genomes was similar [14–17]. Using the Dfam dataset and the RepeatModeler generated dataset; masked bases ranged from 7.86% in *C. angulata* to 10.05% in *C. rhizophorae*, and from 42.85% in *C. gasar* to 47.47% in *C. angulata,* respectively (Table 1—Data Set 1 (Table 4)).

In both genomes, over 90% of proteins had hits in the NR database using Diamond, with 99% being mollusk proteins (Table 1—Data Set 1 (Table 5)). Approximately 80% of hits related to mollusks had query and subject coverage above 90%. Functional annotation with eggNOG identified 34,693 and 26,328 proteins for *C. rhizophorae* and *C. gasar*, respectively (Table 1—Data Set 1 (Table 6)).

These results demonstrate that the draft genomes of *C. gasar* and *C. rhizophorae* represent each species and are sufficiently contiguous to describe genes and repetitive elements, making them suitable references for further research [18, 19]. These data will be used in transcriptome analyses of 3RAD analyses, among other studies (Table 1).

Limitations

Integrating data from different sequencing platforms and individuals posed significant challenges in producing the draft genomes. We explored using Illumina-generated reads alongside PacBio data to form contigs and scaffolds during assembly. Despite trying various methods, we consistently encountered more fragmented assemblies when combining both data types. Therefore, we decided to use Illumina reads to fill gaps within scaffolds generated solely from PacBio reads.

Abbreviations

CLR	Pacific Biosciences Continuous Long Reads
DNA	Deoxyribonucleic Acid
MAS	Marker Assisted Selection
HMW-DNA	High molecular weight DNA
NCBI	National Center for Biotechnology Information
RNA	Ribonucleic Acid
RNA-seq	RNA Sequencing
RPM	Revolutions Per Minute
UFSC	Universidade Federal de Santa Catarina

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-024-01262-6.

Supplementary Material 1.

Authors' contributions

F.H., A.M.S.C., L.M.M.S., R.G.S., F.L.Z., A.C.D.B., C.M.R.M. collected the samples. C.L. and C.M.R.M. identified the samples. F.H. extracted nucleic acids. A.P.C.G. and A.L.G. sequenced the genomes, C.L. performed ONT analyses. L.G.P.A. and N.C.B.L. processed the genomes and the analysis. L.G.P.A. prepared all figures and tables. N.C.B.L., L.G.P.A. and A.T.R.V prepared the manuscript. All authors reviewed the manuscript.

Funding

This study was developed in the frameworks of the Pensa Rio project from Carlos Chagas Filho Foundation for Research Support of the State of Rio de Janeiro (FAPERJ) E-26/010/003027/2014. A.T.R.V. was supported by grants from the National Council for Scientific and Technological Development (CNPq) (307145/2021–2) and FAPERJ (E-26/201.046/2022). C.L. was supported by grants from FAPERJ (210.579/2014). F.H. was supported by grants from CNPq (315816/2021–0) and FAPERJ (E-26/201.458/2021). A.C.D.B. was supported by grants from the CNPq (311725/2021–0). A.M.S.C. was supported by grants from CNPq (303300/2019–1) and FAPERJ (E-26/201.019/2022).

Availability of data and materials

The draft genomes and the raw reads used in this study are publicly available in NCBI, Bioproject accession PRJNA1117898 (https://identifiers.org/ncbi/bioproject:PRJNA1117898). Crassostrea gasar draft genome and Crassostrea rhizophorae draft genome are available at https://identifiers.org/ncbi/nucleotide:JBEEQF000000000.1 [18] and https://identifiers.org/ncbi/nucleotide:JBEOLP000000000.1 [19], respectively.
Tables and Figure are available at: https://doi.org/https://doi.org/10.5281/zenodo.12103998 [20]. Please see Table 1 for details and links to the data.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors consent to this text for publication.

Competing interests

The authors declare no competing interests.

Received: 20 June 2024 Accepted: 21 August 2024
Published online: 03 September 2024

References

1. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.
2. Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, et al. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 2022;40:1332–5.
3. Coombe L, Warren RL, Wong J, Nikolic V, Birol I. ntLink: A Toolkit for De Novo Genome Assembly Scaffolding and Mapping Using Long Reads. Curr Protoc. 2023;3:e733.
4. Coombe L, Li JX, Lo T, Wong J, Nikolic V, Warren RL, et al. LongStitch: high-quality genome assembly correction and scaffolding using long reads. BMC Bioinformatics. 2021;22:534.
5. Nadalin F, Vezzi F, Policriti A. GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics. 2012;13 Suppl 14(Suppl 14):S8.
6. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38:4647–54.
7. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34:i142–50.
8. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013--2015. 2015. http://www.repeatmasker.org.
9. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 2020;117:9451–7.
10. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12:2.
11. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. Methods Mol Biol. 2019;1962:65–95.
12. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38:5825–9.
13. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.
14. Peñaloza C, Gutierrez AP, Eöry L, Wang S, Guo X, Archibald AL, et al. A chromosome-level genome assembly for the Pacific oyster *Crassostrea gigas*. Gigascience. 2021;10:giab020.
15. Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 2012;490:49–54.
16. Qi H, Cong R, Wang Y, Li L, Zhang G. Construction and analysis of the chromosome-level haplotype-resolved genomes of two *Crassostrea* oyster congeners: *Crassostrea angulata* and *Crassostrea gigas*. Gigascience. 2022;12:giad077.
17. Puritz JB, Guo X, Hare M, He Y, Hillier LW, Jin S, et al. A second unveiling: Haplotig masking of the eastern oyster genome improves population-level inference. Mol Ecol Resour. 2024;24:e13801.
18. Genbank. *Crassostrea gasar* draft genome. NCBI. 2024. https://identifiers.org/ncbi/nucleotide:JBEEQF000000000.1. Accessed 15 Jul 2024.
19. Genbank. *Crassostrea rhizophorae* draft genome. NCBI. 2024. https://identifiers.org/ncbi/nucleotide:JBEOLP000000000.1. Accessed 15 Jul 2024.
20. Lima N, Almeida L, Gerber A, Guimarães A, Solé-Cava A, Melo C, et al. The draft genomes of *Crassostrea gasar* and *Crassostrea rhizophorae*: key resources for leveraging oyster cultivation in the Southwest Atlantic. Zenodo; 2024.

Publisher's Note